NTNU
Norwegian University of
Science and Technology

TMA4267 Linear Statistical Models V2014 (15)

MLR: Sums of squares, R^2 and the F-test [3.4-3.6]
NB: different def of SSR and SS used in book

Mette Langaas

To be lectured: February 24, 2014
wiki.math.ntnu.no/emner/tma4267/2014v/start/

Multiple linear regression model with normal assumptions

$$Y_i \quad = \quad \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \text{ for } i = 1, ..., n$$

Matrix formulation:

$$\begin{aligned} \underset{(n\times1)}{\boldsymbol{Y}} &= \underset{(n\times p)}{\boldsymbol{X}}\underset{(p\times1)}{\boldsymbol{\beta}} + \underset{(n\times1)}{\boldsymbol{\varepsilon}} \\ \boldsymbol{\varepsilon} &\sim N_n(\underset{(n\times1)}{\mathbf{0}}, \underset{(n\times1)}{\sigma^2 \boldsymbol{I}}) \end{aligned}$$

where

— β and σ^2 are unknown parameters and

— the design matrix $\boldsymbol{X}$ has ith row $[x_{i1}, x_{i2}, ..., x_{ip}]$.

Sum of Squares - from Chapter 1: Simple linear regression

Let $\hat{y}_i = a + bx_i$.

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SSE}$$

total=regression+error
Cross terms cancel due to the normal equations.
Can we use this also for Multiple LR?
May we write this with matrices?

Partitioning of variance [3.4-3.6]

$$\underset{n\times1}{Y} = \underset{n\times p}{X}\underset{p\times1}{\beta} + \underset{n\times1}{\varepsilon} \qquad \varepsilon \sim N_n(0, \sigma^2 I)$$

Ch1: Sums of squares $\qquad \hat{Y} = X\hat{\beta} = HY$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

SST $\qquad$ SSE $\qquad$ SSR

total $\qquad$ error $\qquad$ regression

Also valid for multiple linear regression, when

$\hat{y}_i = X_i \hat{\beta}$

$\uparrow$ i th row of X

1) Remember SSE $= Y^T(I-H)Y \sim \sigma^2 \chi^2_{n-p}$

2) How to write $\bar{Y} \rightarrow$ the centering matrix.

$\underset{n\times 1}{Y}$ random vector

$$\underset{n\times n}{J} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ \vdots & & & & \\ 1 & \cdots & \cdots & \cdots & 1 \end{bmatrix} = \underset{n\times 1}{1}\,\underset{1\times n}{1^T}$$

$$\underset{n\times 1}{\bar{Y}} = \begin{bmatrix} \bar{Y} \\ \bar{Y} \\ \vdots \\ \bar{Y} \end{bmatrix} = \frac{1}{n}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ \vdots & & & \\ 1 & & & 1 \end{bmatrix}\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \underline{\frac{1}{n} J Y}$$

E4P4: $\frac{1}{n}J$ is symmetric and idempotent with

rank $\left(\frac{1}{n}J\right) = \operatorname{tr}\left(\frac{1}{n}J\right) = 1$

3) $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (Y - \bar{Y})^T (Y - \bar{Y})$

$$= \left(Y - \tfrac{1}{n} J Y\right)^T \left(Y - \tfrac{1}{n} J Y\right) = Y^T \underbrace{\left(I - \tfrac{1}{n} J\right)^T \left(I - \tfrac{1}{n} J\right)}_{\left(I - \frac{1}{n} J\right)} Y$$

$$= Y^T \left(I - \tfrac{1}{n} J\right) Y$$

because $(I - \frac{1}{n} J)$ is also idempotent E4P4

4) What is the distribution of SST?

Theo 3.26 (LIY) $Z \sim N(0, \sigma^2 I)$, A symmetric & idempotent, $\text{rank}(A) = r$ $\Big\}$ $\overset{Y^*}{Z}{}^T \overset{(I-\frac{1}{n}J)}{A} \overset{Y^*}{Z} \sim \sigma^2 \chi^2_r$

and trick $Y^* = Y - X\beta \sim N(0, \sigma^2 I)$

$$(I - \tfrac{1}{n}J)Y^* = (I - \tfrac{1}{n}J)Y - \underbrace{(I - \tfrac{1}{n}J)}X\beta = (I - \tfrac{1}{n}J)Y$$

Difficult : But when $H_0 : \beta_2 = \beta_3 = \dots = \beta_p = 0$

$$\frac{1}{n}JX\beta = \frac{1}{n}J\underbrace{X\begin{bmatrix}\beta_1\\0\\\vdots\\0\end{bmatrix}} = \frac{1}{n}\begin{bmatrix}1 & 1 & 1 & 1\\ \cdot & & & \\ \cdot & & \cdot & 1\end{bmatrix}\begin{bmatrix}\beta_1\\\vdots\\\beta_1\end{bmatrix} = \begin{bmatrix}\beta_1\\\vdots\\\beta_1\end{bmatrix}$$

$$\begin{bmatrix}1 & x_{12} & \cdots & x_{1p}\\ 1 & & & \\ 1 & & & \\ \cdot & & & x_{np}\end{bmatrix}\begin{bmatrix}\beta_1\\0\\\vdots\\0\end{bmatrix} = \begin{bmatrix}\beta_1\\\beta_1\\\vdots\\\beta_1\end{bmatrix} \qquad (I - \tfrac{1}{n}J)X\beta = \begin{bmatrix}\beta_1\\\vdots\\\beta_1\end{bmatrix} - \begin{bmatrix}\beta_1\\\vdots\\\beta_1\end{bmatrix} = 0$$

$$IX\beta = \begin{bmatrix}\beta_1\\\vdots\\\beta_1\end{bmatrix}$$

So

$$Y^T\left(I-\tfrac{1}{n}J\right)Y = Y^{*T}\left(I-\tfrac{1}{n}J\right)Y^* \sim \sigma^2 \chi^2_{n-1}$$

$$\text{rank}\left(I-\tfrac{1}{n}J\right) = tr\left(I-\tfrac{1}{n}J\right) = tr(\underset{n\times n}{I}) - tr(\underset{n\times n}{\tfrac{1}{n}J}) = n-1$$

5) SSR remains

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = (\hat{Y}-\bar{Y})^T(\hat{Y}-\bar{Y})$$

$$\hat{Y}-\bar{Y} = HY - \tfrac{1}{n}JY = \left(H-\tfrac{1}{n}J\right)Y$$

Easy $\left(H-\tfrac{1}{n}J\right)$ is symmetric and idempotent

$$tr\left(H-\tfrac{1}{n}J\right) = tr(H) - tr\left(\tfrac{1}{n}J\right) = p-1$$

Using $Y^* = Y - X\beta$, and assume $H_0: \beta_2 = \beta_3 = \cdots = \beta_p$

Homework!

Thus: $SSR = Y^T\left(H-\tfrac{1}{n}J\right)Y$

$$\sim \sigma^2 \chi^2_{p-1}$$

Sums of squares summary

SSE : always true

$$SSE = \mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbf{Y} \sim \sigma^2\chi^2_{n-p}$$

SSR : under $H_0 : \beta_2 = \beta_2 = \cdots = \beta_p = 0$

$$SSR = \mathbf{Y}^T(\mathbf{H}-\frac{1}{n}\mathbf{J})\mathbf{Y} \sim \sigma^2\chi^2_{p-1}$$

SST : under $H_0 : \beta_2 = \beta_2 = \cdots = \beta_p = 0$

$$SST = \mathbf{Y}^T(\mathbf{I}-\frac{1}{n}\mathbf{J})\mathbf{Y} \sim \sigma^2\chi^2_{n-1}$$

Bingham& Fry (2010) use a slightly different definition of SSR (and also SST), but use the same definition as us for SSE. Our SST is comparing $\mathbf{Y}$ with the mean $\bar{\mathbf{Y}}$, while Bingham & Fry compare $\mathbf{Y}$ with $\mathbf{X}\beta$.

Summing up:

$$SSE = Y^T(I-H)Y \sim \sigma^2 \chi^2_{n-p}$$

under $H_0: \beta_2 = \beta_3 = \dots = \beta_p$

$$SSR = Y^T(H - \tfrac{1}{n}J)Y \sim \sigma^2 \chi^2_{p-1}$$
$$SST = Y^T(I - \tfrac{1}{n}J)Y \sim \sigma^2 \chi^2_{n-1}$$

6) SSR and SSE are independent, using the same reasoning as for $\hat{\beta}$ and SSE → Theo. 4.15 L10

$$AY = (H - \tfrac{1}{n}J)Y \leftarrow SSR$$
$$BY = (I - H)Y \leftarrow SSE$$

$$AB = (H - \tfrac{1}{n}J)(I-H) = H - H^2 - \tfrac{1}{n}J + \underbrace{\tfrac{1}{n}JH}$$

$$= H - H - \tfrac{1}{n}J + \tfrac{1}{n}J = 0 \qquad \tfrac{1}{n}JH = \tfrac{1}{n}J \quad \underline{\text{EYPY}}$$

So AY and BY are independent, and $SSR = Y^TA^TAY$ and $SSE = Y^TB^TBY$ are also independent.

7) $H_0: \beta_2 = \beta_3 = \cdots = \beta_p = 0$

vs H_1: at least one $\beta_i \neq 0$ $i = 2, .., p$

NB: β_1 is intercept, not specified in H_0.

This can be tested by using SSR and SSE.

If H_0 is true $\frac{SSR}{\sigma^2} \sim \chi^2_{p-1}$

and $\frac{SSE}{\sigma^2} \sim \chi^2_{n-p}$ (always), and SSR and SSE are independent

Thus

$$F = \frac{\frac{SSR/\cancel{\sigma^2}}{p-1}}{\frac{SSE/\cancel{\sigma^2}}{n-p}} = \frac{\overset{SSR/(p-1)}{MSR}}{\underset{SSE/(n-p)}{MSE}} \sim F_{p-1, n-p}$$

F-test

$$H_0 : \beta_2 = \beta_2 = \cdots = \beta_p = 0$$

versus at least one of these parameters are not equal 0. (NB: intercept not in hypothesis.)
Remember from ANOVA (chapter 2), we compared SSA with SSE and rejected H_0 when SSA was large compared to SSE? Here we do the same, but we have SSR in place of SSA. Mean square is sums of square divided by degrees of freedom.

$$F = \frac{\frac{SSR/\sigma^2}{p-1}}{\frac{SSE/\sigma^2}{n-p}} = \frac{MSR}{MSE} \sim F_{p-1,n-p}$$

Anova table

Source	SS	df	MS	F	p-val
Regression	SSR	$p-1$	$\frac{SSR}{p-1}$	$\frac{MSR}{MSE}$	
Error	SSE	$n-p$	$\frac{SSE}{n-p}$		
Total	SST	$n-1$			

F-test with acid rain

```
> ds=read.table("http://www.math.ntnu.no/~mettela/TMA4267/
Data/acidrain.txt",header=TRUE)
> dsmat <- as.matrix(ds)
> fit <- lm(dsmat[,1]~dsmat[,2:8])
> anova(fit)
Analysis of Variance Table

Response: dsmat[, 1]
             Df Sum Sq Mean Sq F value    Pr(>F)
dsmat[, 2:8]  7 3.2421 0.46316  34.151 3.904e-09 ***
Residuals    18 0.2441 0.01356
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

R^2

Coefficient of determination: proportion of explained variation

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

same as for the simple linear regression.
R can be seen as the sample correlation coefficient between the observed and fitted data, Y and $\hat{Y}$.
Additional concept: R^2_{adj}:

$$R^2_{adj} = 1 - \frac{MSE}{MST} = 1 - \frac{SSE}{SST} \cdot \frac{n-1}{n-p}$$

if we want to take into account the number of covariates fitted.

Questions

Write down on the piece of paper that you have been given - and return in the box at the entrance.

A Todays quiz: 1=very easy, ..., 5=very difficult. What do you think?

B One word/equation/concept from you to Multiple linear regression mind map.

C Percent of todays lecture that you fully understood was:

D R session (bring your laptop with R installed and go through the steps of analysing a data set with MLR together with the lecturer, and ask any R question you may want).

1. I will join on Wednesday March 5 at 12.15-13.
2. I will join on Friday March 7 at 10.15-11.
3. I will not join.